CS 683 Project 3-Pager & Final Report

Berat Biçer, *21503050*; Batuhan Kaynak, *21501178*; Murat Şahin *22301345*; Mehmet Can Şakiroğlu, *22301343;* Ozan Müjde, *22301347*

Introduction

This project aims to create a user-friendly interface for accessing and utilizing Large Language Models (LLMs) by individuals not well-versed in the technical aspects of Natural Language Processing (NLP). The goal is to bridge the gap between the complexity of LLMs and the accessibility for non-technical users, allowing them to leverage the power of NLP through simple, natural language queries.

Problem Statement

LLMs, while incredibly powerful for various tasks, including NLP, often require technical expertise to be utilized, thus remaining inaccessible to individuals without this technical expertise. Current interfaces to bridge this gap require programming skills and a knowledge of NLP concepts. This limits their user base to a much smaller, specialized, and technical audience. This project seeks to democratize LLM access by providing a simplified and intuitive interface for non-technical users.

Proposed Solution

We propose a serverless, two-layered architecture composed of a front-end service and a cloud-based infrastructure that utilizes multiple services concurrently. The front end lives inside an EC2 instance inside a public subnet accessed via an Internet gateway. This front-end is realized as a web page and is connected to the back-end in a private subnet.

When users arrive, they are authenticated through this web page, which connects to Amazon Cognito via API Gateway. After authentication is complete, users can submit query requests through the front end. This query is paired with a prompt in natural language (NL) that describes the task and the data to be processed. This can be either a text input, or an image; which will be processed in the back-end. The user is also given the opportunity to upload this data in batches, by uploading a folder.

When the user submits a query, this query is passed via API Gateway to a Lambda function. The request is then forwarded to Amazon Textract if the data is composed of images. In this case, its text content is extracted.

After Textract completes execution, the image content and the user prompt are forwarded to the LLM which resides in Amazon Bedrock via a Lambda function. The user prompt is executed in Bedrock and the output is stored in an S3 bucket while the input is also stored in another S3 bucket. Also, the related metadata is stored in an Amazon RDS database. The user is then notified by Amazon SES that the execution is complete via a Lambda function. Then, the web application can query the RDS database via API Gateway with the authorized user's id token to show a user his/her related results from the S3 anytime. Also, the resulting outputs are deleted from the S3 bucket after two days to save space by a Lambda function that is triggered by a CloudWatch event every hour.

Key Features

NLP Task Execution: The interface allows users to request various NLP tasks, such as chatbot interactions, customer support, translation services, etc., using natural language queries.

Authentication and Authorization: Users are required to authenticate themselves before accessing the system. This step enables users to have private storage for their results, ensuring data privacy and security.

Input Methods: Users can provide input texts through natural language queries or images containing text. The system extracts text content from images for processing.

Result Storage: The system can store the text outputs of the users, allowing users to access their results for subsequent processing or analysis.

Key Benefits

Flexible Front-end: A user-friendly front-end interface that resides in an AWS ecosystem allows users to interact with the system, authenticate themselves, and provide input queries. This front-end can be purposefully designed for user experience.

Encryption and Security: Inputs and outputs are transmitted over encrypted channels, secured by Security Groups. The front-end and the back-end are separated where one is in a public subnet while the other is in private. The communication between them is enabled by API Gateway to ensure data security and privacy.

Backend Processing: The backend processes user requests, interacts with the Textract and Bedrock services, stores inputs and outputs of user requests in an S3 Bucket. Furthermore, auxiliary data stored in S3 can be removed post-execution to save space and operational cost.

Scalability: The system has a serverless, scalable architecture that operates on request via Lambda functions to ensure scalability for potential growth in user base.

Monitoring: CloudWatch instances attached to each major component ensures that resource health is monitored 24/7. This allows request throttling, prevents malicious or excessive usage, allows smart resource management and can save operational costs.

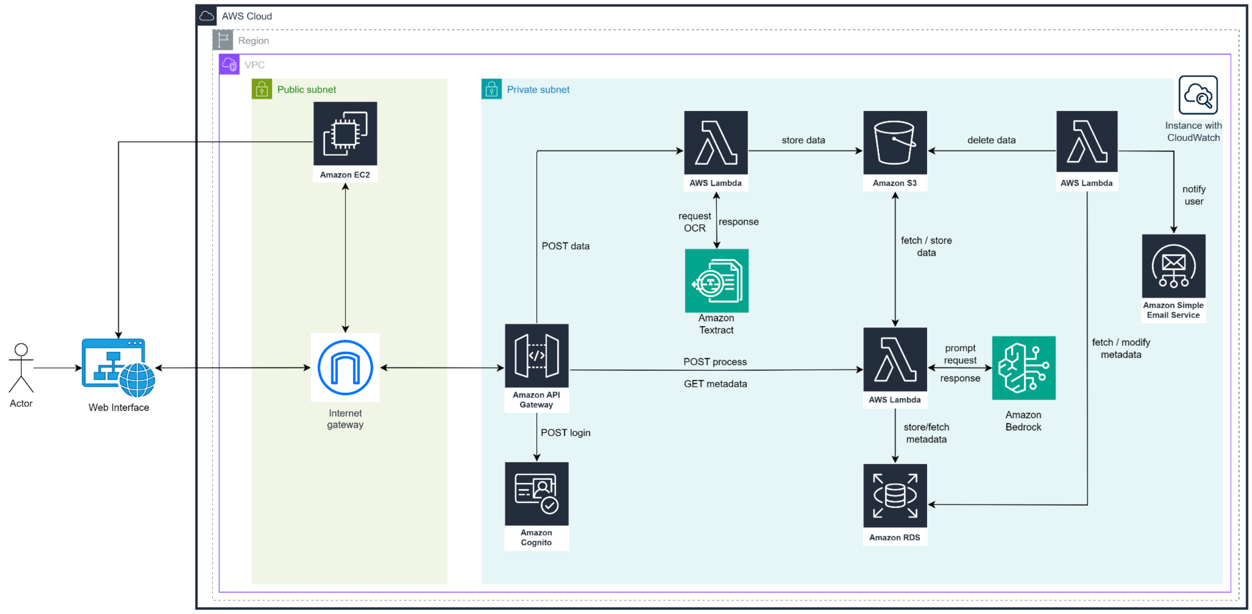


Figure 1: Proposed Architecture

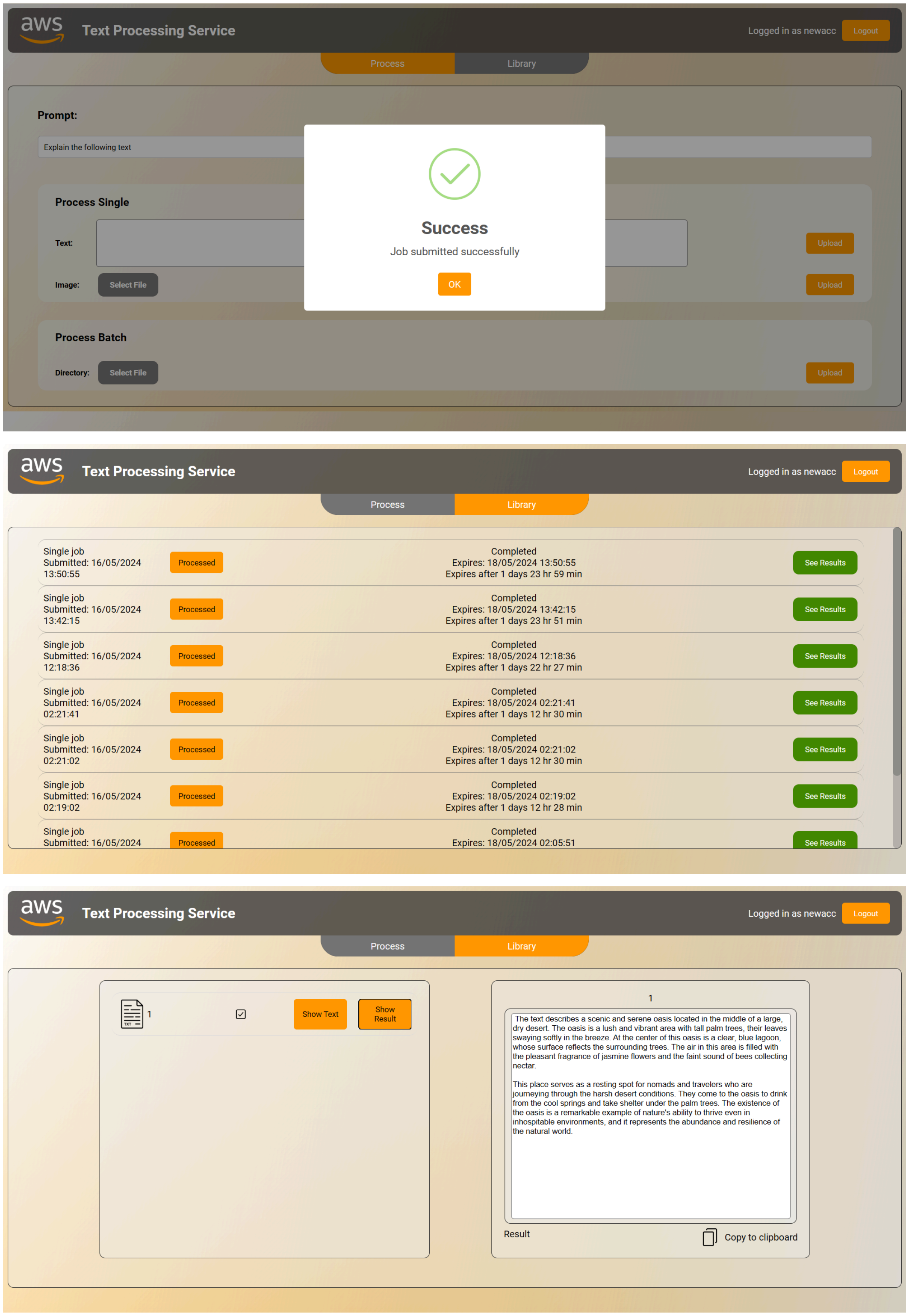
Example Use Case

Consider that a user wishes to run a profanity filter on messages that are embedded in images. The user interacts with our dedicated front-end to provide these inputs through encrypted channels. Upon the request's arrival via API Gateway, we perform a secure login via Cognito. The image is then forwarded to Textract where the embedded text within is obtained. Afterwards, the user prompt and the embedded text are given to the LLM that resides in Bedrock. LLM executes the user-specified task. Upon completion, the response is pushed to S3 alongside with the input to another S3. We store the related metadata in an RDS Database instance for future business analytics, notify the user via Amazon SES that the output is ready, and return the output to the user via the dedicated front-end.

Conclusion

This project aims to make NLP tasks accessible to non-technical users by providing a user-friendly interface. The system's authentication, input methods, and result storage capabilities ensure a secure and efficient user experience. The system's key features include executing various NLP tasks through natural language queries, authentication for private and secure execution, input options through text or image extraction, and mass storage of text outputs. It offers a flexible front-end for user interaction, encryption for security, and efficient processing with Textract and Bedrock services. The system is highly scalable due to its serverless architecture and employs CloudWatch for continuous monitoring and smart resource management.

Extra (Views From the User Interface)



Prompt:

Explain the following text

Process Single

Text:

Image:

Process Batch

Directory:

Success

Job submitted successfully

Single job
Submitted: 16/05/2024
13:50:55
Processed
Completed
Expires: 18/05/2024 13:50:55
Expires after 1 days 23 hr 59 min
See Results

Single job
Submitted: 16/05/2024
13:42:15
Processed
Completed
Expires: 18/05/2024 13:42:15
Expires after 1 days 23 hr 51 min
See Results

Single job
Submitted: 16/05/2024
12:18:36
Processed
Completed
Expires: 18/05/2024 12:18:36
Expires after 1 days 22 hr 27 min
See Results

Single job
Submitted: 16/05/2024
02:21:41
Processed
Completed
Expires: 18/05/2024 02:21:41
Expires after 1 days 12 hr 30 min
See Results

Single job
Submitted: 16/05/2024
02:21:02
Processed
Completed
Expires: 18/05/2024 02:21:02
Expires after 1 days 12 hr 30 min
See Results

Single job
Submitted: 16/05/2024
02:19:02
Processed
Completed
Expires: 18/05/2024 02:19:02
Expires after 1 days 12 hr 28 min
See Results

Single job
Submitted: 16/05/2024
Processed
Completed
Expires: 18/05/2024 02:05:51
See Results

The text describes a scenic and serene oasis located in the middle of a large, dry desert. The oasis is a lush and vibrant area with tall palm trees, their leaves swaying softly in the breeze. At the center of this oasis is a clear, blue lagoon, whose surface reflects the surrounding trees. The air in this area is filled with the pleasant fragrance of jasmine flowers and the faint sound of bees collecting nectar.

This place serves as a resting spot for nomads and travelers who are journeying through the harsh desert conditions. They come to the oasis to drink from the cool springs and take shelter under the palm trees. The existence of the oasis is a remarkable example of nature's ability to thrive even in inhospitable environments, and it represents the abundance and resilience of the natural world.

Figure 2: Sample Views From Front-End

aws pricing calculator

Contact your AWS representative: Contact Sales

Export date: **4/21/2024**

Language: **English**

Estimate URL: **https://calculator.aws/#/estimate?id=092fc9893a71844c9630a869fe6c98010d36e2df**

Estimate summary

Upfront cost	Monthly cost	Total 12 months cost
5.66 USD	**210.24 USD**	**2,528.54 USD** Includes upfront cost

Detailed Estimate

Name	Group	Region	Upfront cost	Monthly cost
AWS Lambda	No group applied	EU (Frankfurt)	0.00 USD	0.00 USD

Status: -
Description:
Config summary: Architecture (x86), Architecture (x86), Invoke Mode (Buffered), Amount of ephemeral storage allocated (512 MB), Number of requests (100000 per month)

Amazon Simple Storage Service (S3)	No group applied	EU (Frankfurt)	5.66 USD	0.45 USD

Status: -
Description:
Config summary: S3 Standard storage (16 GB per month), S3 Standard Average Object Size (16 KB), PUT, COPY, POST, LIST requests to S3 Standard (1024), GET, SELECT, and all other requests from S3 Standard (1024), Data returned by S3 Select (16 GB per month), Data scanned by S3 Select (16 GB per month) DT Inbound: Internet (20 GB per month), DT Outbound: Not selected (0 TB per month)

Amazon Cognito	No group applied	EU (Frankfurt)	0.00 USD	0.00 USD

Status: -
Description:
Config summary: Advanced security features (Disabled), Number of monthly active users (MAU) (1024)

Amazon Textract	No group applied	EU (Frankfurt)	0.00 USD	36.86 USD

Status: -
Description:
Config summary: Number of pages (24576)

Amazon RDS for PostgreSQL	No group applied	EU (Frankfurt)	0.00 USD	41.75 USD

Status: -
Description:
Config summary: Storage volume (General Purpose SSD (gp2)), Storage amount (1 GB), Nodes (1), Instance Type (db.t3.micro), Utilization (On-Demand only) (100 %Utilized/Month), Deployment Option (Single-AZ), Pricing Model (OnDemand)

Amazon Bedrock	No group applied	EU (Frankfurt)	0.00 USD	82.00 USD

Status: -
Description:
Config summary: Number of Input tokens (30 million per month), Number of output tokens (20 million per month)

Amazon Simple Notification Service (SNS)	No group applied	EU (Frankfurt)	0.00 USD	19.98 USD

Status: -
Description:
Config summary: EMAIL/EMAIL-JSON Notifications (1 million per month)

Amazon CloudWatch	No group applied	EU (Frankfurt)	0.00 USD	5.75 USD

Status: -
Description:
Config summary: Number of Metrics (includes detailed and custom metrics) (10), GetMetricData: Number of metrics requested (10000), Number of other API requests (10000), Number of Lambda functions (1), Number of requests per function (5 per minute)

Amazon Virtual Private Cloud (VPC)	No group applied	EU (Frankfurt)	0.00 USD	4.85 USD

Status: -
Description:
Config summary: Number of In-use public IPv4 addresses (1) DT Inbound: Internet (16 GB per month), DT Outbound: Internet (8 GB per month), DT Intra-Region: (24 GB per month), Data transfer cost (1.2)

Amazon EC2	No group applied	EU (Frankfurt)	0.00 USD	3.80 USD

Status: -

Description:

Config summary: Tenancy (Shared Instances), Operating system (Linux), Workload (Consistent, Number of instances: 1), Advance EC2 instance (t3.micro), Pricing strategy (EC2 Instance Savings Plans 3yr No Upfront), Enable monitoring (disabled), DT Inbound: Not selected (0 TB per month), DT Outbound: Not selected (0 TB per month), DT Intra-Region: (0 TB per month)

Amazon API Gateway	No group applied	EU (Frankfurt)	0.00 USD	14.80 USD

Status: -

Description:

Config summary: REST API request units (millions), Cache memory size (GB) (None), WebSocket message units (thousands), HTTP API requests units (millions), Average size of each request (34 KB), Average message size (32 KB), Requests (4 per month)

Acknowledgement

AWS Pricing Calculator provides only an estimate of your AWS fees and doesn't include any taxes that might apply. Your actual fees depend on a variety of factors, including your actual usage of AWS services. Learn more